Vision Research 178 (2021) 28–40

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Exploring biological motion perception in two-stream convolutional neural networks

Yujia Peng[a,*], Hannah Lee[a], Tianmin Shu[b,c], Hongjing Lu[a,b]

[a] Department of Psychology, University of California, Los Angeles, United States
[b] Department of Statistics, University of California, Los Angeles, United States
[c] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States

ARTICLE INFO

Keywords:
Biological motion
Action recognition
Two-stream convolutional neural network
Local image motion
Inversion effect
Motion congruency
Causal perception

ABSTRACT

Visual recognition of biological motion recruits form and motion processes supported by both dorsal and ventral pathways. This neural architecture inspired the two-stream convolutional neural network (CNN) model, which includes a spatial CNN to process appearance information in a sequence of image frames, a temporal CNN to process optical flow information, and a fusion network to integrate the features extracted by the two CNNs and make final decisions about action recognition. In five simulations, we compared the CNN model's performance with classical findings in biological motion perception. The CNNs trained with raw RGB action videos showed weak performance in recognizing point-light actions. Additional transfer training with actions shown in other display formats (e.g., skeletal) was necessary for CNNs to recognize point-light actions. The CNN models exhibited largely viewpoint-dependent recognition of actions, with a limited ability to generalize to viewpoints close to the training views. The CNNs predicted the inversion effect in the presence of global body configuration, but failed to predict the inversion effect driven solely by local motion signals. The CNNs provided a qualitative account of some behavioral results observed in human biological motion perception for fine discrimination tasks with noisy inputs, such as point-light actions with disrupted local motion signals, and walking actions with temporally misaligned motion cues. However, these successes are limited by the CNNs' lack of adaptive integration for form and motion processes, and failure to incorporate specialized mechanisms (e.g., a life detector) as well as top-down influences on biological motion perception.

1. Introduction

One of the most sophisticated abilities supported by the human visual system is the recognition of human body movements. In daily life, humans can readily recognize actions despite changes in body forms and appearance (e.g., different costumes and clothing texture, viewpoints, and occlusions). Even for highly impoverished and rarely observed stimuli such as point-light displays (Johansson, 1973), in which a few disconnected dots depict joint movements, the human visual system can still recognize actions despite visual noise (Neri, Morrone, & Burr, 1998; Lu, 2010). In addition to action recognition, humans can identify other characteristics of point-light actors, including gender (Kozlowski & Cutting, 1977; Pollick, Kay, Heim, & Stringer, 2005), identity (Cutting & Kozlowski, 1977; Pavlova, 2011), personalities (e.g., Brownlow, Dixon, Egbert, & Radcliffe, 1997), emotions (Dittrich, Troscianko, Lea, & Morgan, 1996), social interactions (Thurman & Lu, 2014), and causal intention (Peng, Thurman, & Lu, 2017).

Over several decades, psychophysical and neuroscience research has advanced our understanding of the underlying processes and mechanisms supporting the robust perception of biological motion. Early work hypothesized that point-light actions are analyzed primarily in the dorsal (motion) pathway, with recognition achieved by spatiotemporal integration of motion information specific to body movements (Mather, Radford, & West, 1992). However, this view was challenged by neuropsychological studies showing that patients with lesions in the dorsal pathway (i.e., V5/MT) maintain the ability to recognize actions in point-light displays (Vaina, Lemay, Bienfang, Choi, & Nakayama, 1990). Psychophysical studies provided further evidence that human observers have no trouble recognizing point-light actions with degraded or perturbed local motion (Beintema & Lappe, 2002; van Boxtel & Lu, 2015), or when point-light actions are embedded within a cloud of noise dots with the same joint motion trajectories (e.g., Cutting, Moore, & Morrison, 1988).

* Corresponding author.
E-mail addresses: yjpeng@ucla.edu (Y. Peng), leehannah@ucla.edu (H. Lee), tshu@mit.edu (T. Shu), hongjing@ucla.edu (H. Lu).

https://doi.org/10.1016/j.visres.2020.09.005
Received 27 October 2019; Received in revised form 29 May 2020; Accepted 15 September 2020

These findings suggest that biological motion perception does not entirely rely on the dorsal pathway, or motion processing alone. In fact, bodily form and appearance information have been found to also play important roles in the perception of biological motion (Lange, Georg, & Lappe, 2006). For example, Pinto and Shiffrar (1999), showed that violation of the hierarchical structure of body form can significantly disrupt the detection of biological motion. Lu (2010) showed that when body structural information was eliminated but local motion information was intact in the stimuli, human observers failed to discriminate walking directions in biological movement, suggesting the necessity of structural information for refined discrimination in biological motion. Theusner, de Lussanet, and Lappe (2011) found that adaptation to biological motion elicits both form aftereffects and motion aftereffects, suggesting the co-existence of form processes and motion processes in analyzing biological motion information. In addition, fMRI experiments have shown that biological motion is processed by both ventral and dorsal pathways in the brain. Point-light displays not only activate the dorsal stream involving the motion selective regions such as MT/MST, but also the ventral stream with a projection from primary visual cortex to inferotemporal cortex that processes object appearance information (Grossman & Blake, 2002). Finally, numerous studies have established that a region selective for biological motion, posterior superior temporal sulcus (STSp), integrates motion processing and appearance processing carried out by two separate pathways (Grossman et al., 2000; Vaina, Solomon, Chowdhury, Sinha, & Belliveau, 2001; Bonda, Petrides, Ostry, & Evans, 1996; Thurman, van Boxtel, Monti, Chiang, & Lu, 2016).

Inspired by the two-stream processing of biological motion perception in the brain, Giese and Poggio (2003) developed a computational model with two parallel processing streams: a ventral pathway and a dorsal pathway. The ventral pathway is specialized for the analysis of body forms in static image frames. The dorsal pathway is specialized for processing optic-flow/motion information. Both pathways comprise a hierarchy of feature detectors with increasing receptive fields and increasing complexity in encoding form or motion patterns. In the computational model, the ventral pathway starts from the first layer, which consists of local orientation detectors with small receptive fields that approximate neurons in the primary visual cortex (Dow, Snyder, Vautin, & Bauer, 1981). The second layer contains position- and scale-invariant bar detectors, corresponding to position-invariant cells in visual areas V1, V2, and V4 (Hegdé & Van Essen, 2000; Gallant, Braun, & Van Essen, 1993). The third layer consists of snapshot neurons selective to body shapes for form processing, simulating neurons in inferotemporal cortex (area IT) that are selective for complex shapes (Logothetis, Pauls, & Poggio, 1995). In the dorsal pathway, the first layer consists of local motion detectors that correspond to direction-selective neurons in V1 and MT (Rodman & Albright, 1989). The second layer simulates neurons in MT (Smith & Snowden, 1994) with larger receptive fields that are sensitive to optical flow information based on spatial integration of local motions. The third layer contains optical flow pattern neurons that are selective for complex movement patterns, simulating neurons in STS (Oram & Perrett, 1994).

The computational model described by Giese and Poggio (2003) provided a parsimonious framework for biological motion perception. The model, developed to incorporate experimental and biological constraints, can account for many empirical findings in psychophysical experiments using point-light displays and remains one of the most influential computational models in the field. It should be noted that many filter parameters used in the model are either adapted from neurophysiological measures or manually tuned. Although these parameters provide a connection between modeling and neural activities, it remains unclear whether these parameters in the network can be learned from natural statistics with a large number of action videos.

In recent years, with the rise of deep learning models, large-scale networks can be trained with millions of videos to recognize human actions in natural scenes. The significant advances began with a two-stream model developed by Simonyan and Zisserman (2014). The two-stream model extends deep convolutional neural networks (CNNs) (LeCun, Bottou, Bengio, & Haffner, 1998; Krizhevsky, Sutskever, & Hinton, 2012) for action recognition to include two CNNs: a spatial CNN that takes pixel-level intensity as the input, and a temporal CNN that takes optical flow as the input. Thus, a spatial CNN processes appearance information and is trained to perform action recognition from a sequence of static image frames, and a temporal CNN processes optical flow between image frames and is trained to recognize actions from motion information. Each stream in the model adopts a CNN architecture, and the features extracted from the two streams are combined via a late fusion network to obtain the final recognition decision. The two-stream model performed well on action classification for two challenging datasets: UCF-101, which includes 13,320 videos covering 101 action types (Soomro, Zamir, & Shah, 2012), and HMDB-51, which includes 6766 videos covering 51 action types (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011). The two-stream CNNs achieved accuracy levels of 88% for the UCF101 dataset (compared to the chance level performance of 1%), and 59.4% for the HMDB-51 dataset (compared to the chance level of 2%). An improved version of the two-stream model further increased its accuracy to 92.5% for UCF101 and 65.4% for HMDB51 (Feichtenhofer, Pinz, & Zisserman, 2016).

Given that the two-stream CNN model exhibits close-to-human performance in recognizing actions from raw videos, and uses an architecture similar to the brain pathways involved in biological motion perception, this model provides an opportunity to examine how well a deep learning model trained with big data can account for human performance in classic psychophysical experiments on biological motion perception, and to gauge how different processing pathways contribute to the final decisions for various action recognition tasks. The present paper reports a series of such tests. In Simulation 1 we tested whether the two-stream CNN can recognize point-light actions after training with natural RGB videos. We also explored whether additional transfer training with skeletal displays can enable the model to recognize actions from point-light displays. In Simulation 2 we examined whether the two-stream CNN model exhibits some degree of viewpoint-invariant recognition for biological motion. Simulation 3 investigated whether the model exhibits inversion effects as are observed for humans in biological motion perception across a range of experimental conditions (Troje & Westhoff, 2006). Simulation 4 tested whether the two-stream CNN model can recognize actions in noisy displays, such as sequential position point-light displays (Beintema & Lappe, 2002). Simulation 5 examined the performance of the two-stream CNN model in refined discrimination for different types of walking stimuli, including intact forward walking, backward walking, moonwalk, and in-place walking. Additionally, we tested whether the two-stream CNN model can be trained to discriminate between action with and without motion congruency and whether the model shows sensitivity to causal relations underlying motion congruency.

2. Model structure and training for action recognition

2.1. Model architectures of CNNs

The two-stream CNN model relies on processing two types of information to classify a video into alternative action categories. One source of information is the pixel-level appearance of moving body in a sequence of static images, and the other is motion (usually represented by optical flow fields, i.e., the spatial displacement of each pixel between adjacent frames; Horn & Schunck, 1981). This two-stream architecture is consistent with neurophysiological evidence that action processing involves both ventral and dorsal pathways in the brain, and integrates the information at action sensitive regions in the temporal lobe. The two-stream architecture is also consistent with the computational framework proposed in the biological motion literature (Giese & Poggio, 2003).

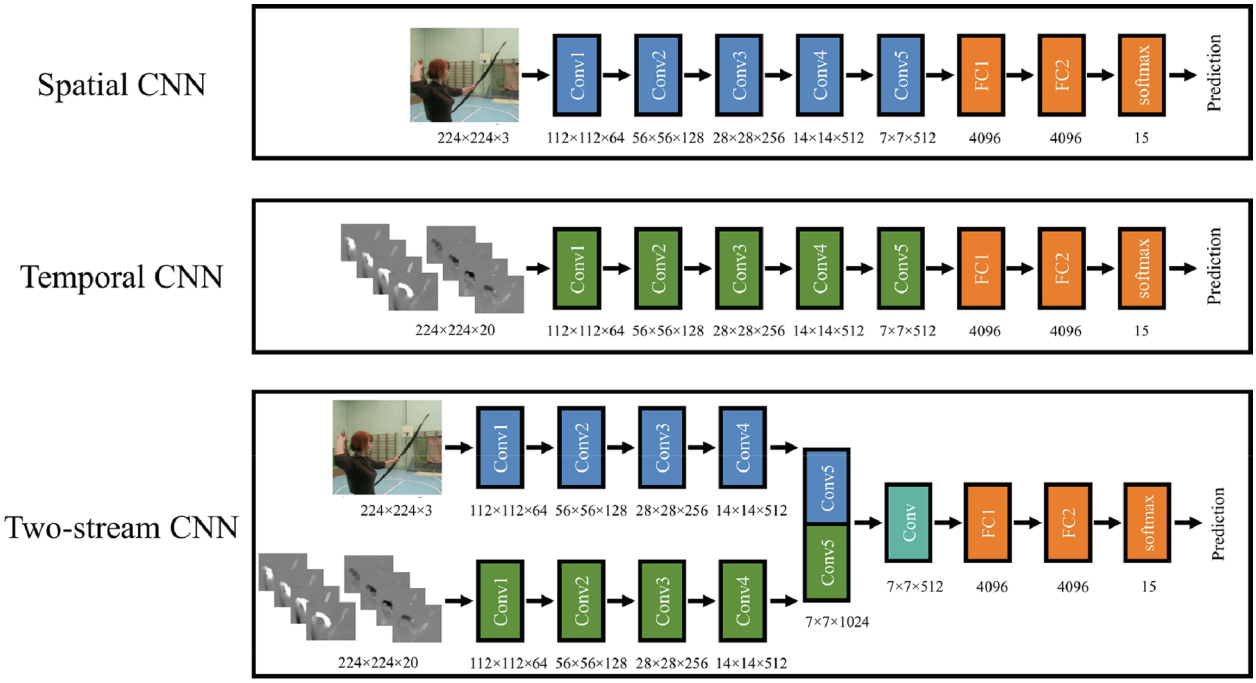

Fig. 1. The architectures of the spatial CNN (top), the temporal CNN (middle), and the two-stream CNN (bottom) using the VGG16 networks. Conv: convolutional layer; FC: fully connected layer. The spatial and temporal CNNs each have 5 convolutional layers and 3 fully connected layers. The convolutional 5 layers were fed into the fusion network with one convolutional layer and 3 fully connected layers.

Fig. 1 (top and middle) shows two single-stream CNN models using the architecture including five convolutional layers for feature extraction, followed by two fully-connected (FC) layers to process either appearance information or optical flow information for action recognition. The spatial CNN for processing the appearance information takes the input of the three channels of an RGB image. The temporal CNN for processing the motion information takes a stack of optical flow vector fields spanning a few consecutive frames (we use 10 frames for all simulations) as the input. In the present paper, we use the spatial CNN and the temporal CNN to model recognition performance from each distinctive stream, corresponding to the spatial pathway and the motion pathway, respectively.

As shown in Fig. 1 bottom panel, a two-stream CNN model combines the spatial process and the motion process to achieve fusion of decisions. The basic approach to construct a two-stream CNN is to take the outputs of one layer in the spatial CNN and the outputs of one layer in the temporal CNN, and concatenate the activities as the joint inputs to an additional fusion network (usually a few FC layers) that perform final action recognition. Simulation work (Feichtenhofer et al., 2016) suggests that fusion of the activities in the final convolutional layers (i.e., "conv5") of both streams consistently yields the best recognition accuracy across different datasets. Accordingly, the present paper adopted this fusion architecture shown in Fig. 1 (bottom) for the two-stream CNN model. Specifically, the two-stream CNN model uses the fusion layer to first stack the outputs from the "conv5" layers of spatial CNN and temporal CNN. The stacked activities from $7 \times 7 \times 1024$ tensors provide inputs to a convolutional layer consisting of 512 filters, followed by three FC layers including a softmax decision layer.

2.2. Model training with natural action videos

The present paper used the Human 3.6 M dataset (Ionescu, Li, & Sminchisescu, 2011; Ionescu, Papava, Olaru, & Sminchisescu, 2014) to train the CNN models with raw RGB videos. The Human 3.6 M dataset (http://vision.imar.ro/human3.6m/description.php) includes a total of 15 categories of actions: giving directions, discussing something with someone, eating, greeting someone, phoning, posing, purchasing (i.e., hauling up), sitting, sitting down, smoking, taking photos, waiting, walking, walking dog, and walking together. Each action was performed twice by each of the seven actors. This dataset provides both raw RGB videos and motion capture data including joint 3D coordinates that can be used to generate skeletal and point-light displays of the actions from different viewpoints.

The CNN models are trained to perform an action classification task with the 15 categories defined in the Human 3.6 M dataset. The action category with the highest score in the softmax layer is considered to be the model prediction for that instance. We follow a two-phase protocol to train the network as developed by Feichtenhofer et al. (2016). We first train the single-stream networks (spatial CNN and temporal CNN) independently with the task of 15-category action recognition. Then activities from the conv5 layers of these two trained single-stream CNNs are concatenated as inputs to train the fusion network in the two-stream CNN. Simulation codes used in the current study are available online (https://github.com/yjpeng11/BM_CNN).

In the following simulations, we used transfer training methods to fine-tune the CNNs to enhance the generalization of the pre-trained CNNs, so that these models can perform with new action stimuli and visual tasks. Transfer training uses a small set of training data and to adopt the previously learned features to new visual tasks. This technique has shown success in extending pre-trained networks to perform in psychophysical tasks, such as shape recognition and object recognition (e.g., Baker, Lu, Erlikhman, & Kellman, 2018). Specifically, two types of transfer training were applied for different simulations in the present paper: *unrestricted transfer training* to adjust all connection weights to optimize CNN model performance. Simulation 1, 2, and 3 used the unrestricted transfer training when a large number of training data is available. *Restricted transfer training*, where the number of classes is changed in the decision layer and learning is limited to the connections between dense FC layers and the decision layer. Simulation 3, 4, and 5 used the restricted transfer training when relatively a small set of training data is available.

The training of CNN models was assigned with a maximum of 100 epochs. Each epoch ran through a series of mini-batches of size 16. Gradient descent was calculated after each mini-batch through an SGD

optimizer (learning rate 10^{-4}) to update model weights. After each epoch, validation loss was calculated and model weights were saved if validation loss decreased compared to the previous epoch. Training will terminate before reaching 100 epochs if validation loss remains without an increase for 10 consecutive epochs. Drop-out operations were implemented for training FC layers with a fraction of the input units to drop of 0.5 to prevent overfitting.

3. Simulation 1: Recognition of point-light actions

The hallmark demonstration of human biological motion perception is that people can recognize actions from sparsely disconnected points, even though such point-light displays are rarely observed in natural environments. Ever since Johansson (1973), numerous studies have shown that humans can recognize point-light actions without previous exposure. Simulation 1 first trained the two-stream CNN model using raw videos, and then tested how well the model can recognize actions in point-light displays. If the two-stream CNN model exhibits a certain degree of robustness in action recognition, as do humans, this would demonstrate some ability to recognize point-light actions after training with natural videos of human actions. If the model trained with natural RGB videos fails to generalize to point-light displays, we will follow-up with transfer learning to explore the possibility that the model's generalization ability can be enhanced using a more diverse set of training stimuli.

3.1. Stimuli and procedure

To train and test the CNNs, a large number of RGB videos were generated. The Human 3.6 M dataset included 210 actions each lasting for 1 to 2 min and performed by 7 actors. Actions were recorded from 4 different viewpoints simultaneously. Actions were segmented to a set of short 5 s clips with a non-overlapping temporal window from the beginning of each action to the end. This temporal segmentation procedure yielded 7962 videos each of 5 s duration. Each video contained 150 frames with a 30 fps sampling rate. The image resolution of videos was 1000 by 1000 pixels. 80% of the original 7962 videos were randomly chosen for training and the rest 20% for cross-validation. To enable the CNNs to acquire position invariance for recognition, variants of the raw videos were included in the training by imposing image transformations. For each original RGB video, 5 additional versions were generated by altering image scale and position of the actors, including a zoom-in version with scale enlarged by a factor of 1.67; and spatially-shifted versions in which the human figure was shifted toward the top-left, top-right, bottom-left, and bottom-right corners, with scale enlarged by a factor of 1.25. Including the variants of RGB videos, we used a total of 6370*6 videos for training and the remaining 1592*6 videos for cross-validation testing. For the spatial CNN model, video frames were down-sampled so that one out of every 10 frames were provided as inputs. For the temporal CNN model, optical flow information calculated from every consecutive 10 frames were provided as inputs.

As the Human 3.6 M dataset provides motion capture data with 3D joint coordinates, we were able to generate skeletal and point-light displays using the tracked joint positions for the same sets of actions. A total of motion capture data from the 1976 actions were used to generate skeleton and point-light videos from any viewpoint. Point-light videos were generated with 13 dots on major joints of an actor: head point, two points on shoulders, elbows, hands, hips, knees, and feet. Sample frames of videos from a subset of action categories are shown in Fig. 2. Skeletal and point-light displays were generated using the Bio-Motion toolbox (van Boxtel & Lu, 2013).

3.2. Results and discussion

Table 1 shows the recognition performance for the validation set after training with raw RGB videos, and corresponding testing accuracy for skeleton and point-light displays. After training with RGB videos, the model achieved good recognition performance for recognizing actions presented in the raw video format. The spatial CNN based on appearance and the two-stream CNN with fusion network yielded better recognition performance (> 0.85) than did the temporal CNN based on optimal flow information (0.70).

To test whether the CNNs can generalize to other display formats of actions, we used the test set of actions in skeletal and point-light displays. The results showed that the three CNNs have limited ability in generalizing action recognition to untrained formats of displays in which visual information, especially appearance of actors, was significantly reduced. As shown in Table 1, the three CNN models showed poor recognition performance for the skeletal displays (0.07, 0.15, and 0.11, respectively) and for point-light actions (0.09, 0.18, and 0.16). Although all CNN models except the spatial CNN yielded accuracy higher than the chance level of 0.067 (i.e., one out of 15 categories), the significant reduction of performance for CNN models in recognizing point-light actions was much worse than is observed for humans. Interestingly, the temporal CNN processing motion information showed a slightly higher accuracy (0.18) in recognizing point-light actions than did either the spatial CNN based on appearance information (0.08). This result is consistent with previous modeling work showing that spontaneous generalization from natural action stimuli to point-light displays is more robustly supported by the motion pathway than by the form pathway (Giese & Poggio, 2003).

As recognition performance for point-light actions was low for CNNs trained with RGB natural videos, we introduced additional unrestricted transfer training to enable the CNNs to perform well with the recognition task with skeletal displays. The parameters (i.e., connection weights) in CNNs trained with RGB videos were used as initial values in retraining the CNN models with skeleton videos with the same two-phase protocol. For skeletal displays, 7 different viewpoints were generated for each action, ranging from 30° counter-clockwise from the central viewpoint to 30° clockwise from the central viewpoint, with a step size of 10°. The image size of skeletal human figures was controlled to be roughly the same size as actors in natural videos. A total of 13,769 skeleton videos (1967 actions * 7 viewpoints) were generated, of which 80% (i.e., 11,015) were used for training and the remaining 20% (2754) for validation. After transfer training, 1967 frontal viewpoint point-light actions were used to compute recognition accuracy for testing.

This transfer training with skeletal actions enabled the CNN models to succeed in recognizing actions in skeletal displays, showing high accuracy in recognizing actions in the validation testing set 0.99 for the spatial CNN processing appearance information, 0.98 for temporal CNN processing motion information, and 0.99 for the two-stream CNN with fusion network. When tested with point-light displays (see Table 2), the temporal CNN based on motion processing yielded an accuracy of 0.42 in recognizing actions, significantly higher than chance (0.067 for classifying 15 categories). The spatial CNN based on appearance processing yielded low recognition performance (0.24), although significantly above chance. These results provide converging evidence that motion processing plays a primary role in recognizing point-light actions, with form processing serving as a secondary process that also contributes to the recognition of point-light actions (Johansson, 1973; Beintema & Lappe, 2002; Giese & Poggio, 2003; Lu, 2010). The two-stream CNN with fusion network achieved an accuracy of 0.23 in recognizing point-light actions. Although this recognition performance was above chance level, test performance for the fusion network was worse than that for the single-pathway spatial CNN and temporal CNN, suggesting that the fusion network may adopt suboptimal integration of the two pathways for recognizing point-light actions and demonstrate limited generalization ability. Supplemental section 1 includes confusion matrices of recognition judgments for three CNN models.

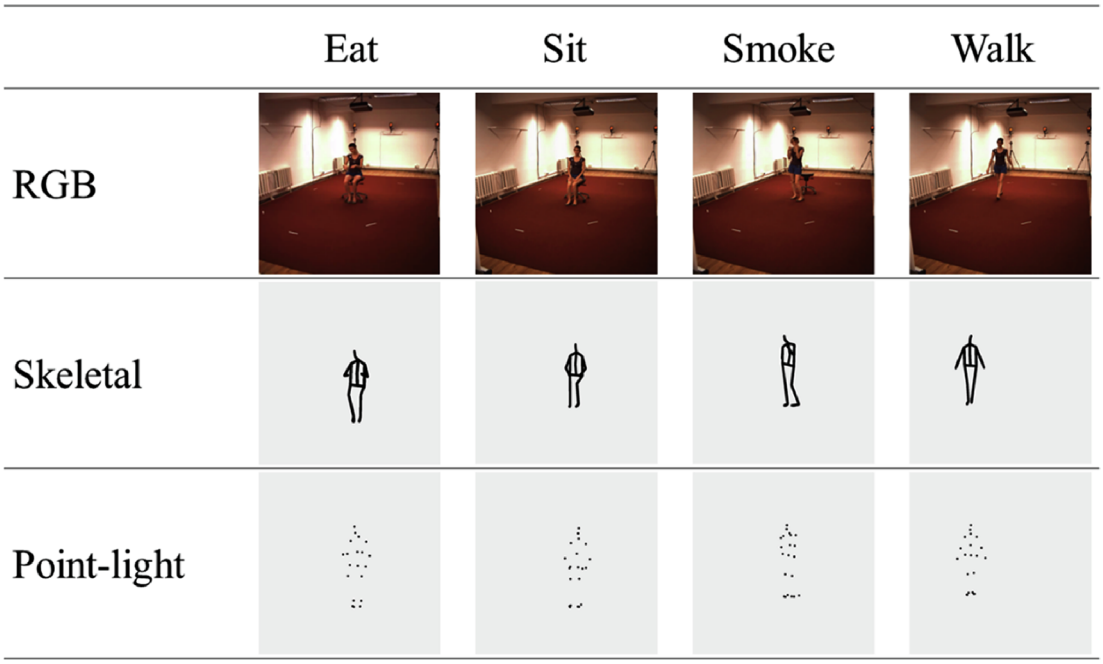

Fig. 2. Sample video frames in the RGB display (top), the skeletal display (middle), and the point-light display (bottom). Videos were taken from the of Human 3.6 M dataset and generated from corresponding motion capture data.

Table 1
Training validation and testing accuracy of action classification after training the CNNs using RGB videos. The CNN models are trained with RGB videos, and tested with skeleton and point-light displays. The chance-level accuracy is 0.067 (i.e., one out of 15 categories).

	Training validation RGB videos	Testing 1 Skeleton displays	Testing 2 Point-light displays
Appearance (spatial CNN)	0.85	0.07	0.08
Motion (temporal CNN)	0.70	0.15	0.18
Fusion (spatial + temporal two-stream CNN)	0.87	0.11	0.16

Table 2
Training and testing accuracy of action classification in Simulation 1, in which training used RGB videos and skeletal displays, and testing used point-light displays.

	Transfer training validation Skeleton displays	Testing Point-light displays
Appearance (spatial CNN)	0.99	0.24
Motion (temporal CNN)	0.98	0.42
Fusion (spatial + temporal two-stream CNN)	0.99	0.23

4. Simulation 2: Viewpoint-dependent effects in action recognition

In the domain of biological motion, researchers have observed viewpoint-dependent performance in identity recognition (Troje, Westhoff, & Lavrov, 2005; Jokisch, Daum, & Troje, 2006) and gender classification from walking gaits (Mather & Murdoch, 1994; Troje, 2002), as people show better performance in frontal view than in profile view for point-light displays. A recent MEG study (Isik, Tacchetti, & Poggio, 2017) found evidence that both viewpoint-dependent representations and viewpoint-invariant representations are used in action recognition for point-light displays, as brain activities can be decoded for both within-view and cross-view recognition, but at different time points. In Simulation 2, we examined whether the two-stream CNN model exhibits viewpoint-dependent effects in action recognition.

4.1. Stimuli and model training

A two-step training procedure similar to that employed in Simulation 1 was used in this simulation. First, the CNN models were trained using Human 3.6 M videos to recognize the 15 categories of actions, with 80% of raw RGB video instances used for training and the remaining 20% used as the validation set to test the model's performance. After reaching a saturated accuracy for the validation set, the trained parameters for the CNN models were saved as initial values for the subsequent transfer training. Second, additional unrestricted transfer training was conducted using 1967 skeleton videos showing only a frontal viewpoint (i.e., the model never saw skeleton videos from other viewpoints), with 80% being used for training and the rest 20% for validation. In Simulation 2, the testing stimuli were 7968 skeleton videos in ± 30° view and ± 90° (profile) view, rotated either clockwise or counterclockwise from the frontal view.

4.2. Results and discussion

As shown in Fig. 3, recognition accuracy for frontal view actions was in the range of 0.72–0.81 (with a chance-level performance of 0.067 across 15 categories) for the three CNN models, indicating the success of transfer training with skeletal displays for action recognition given the relatively small number of training instances. Testing accuracy for 30° views showed slightly decreased but still relatively high recognition accuracy in the range of 0.65–0.73 for spatial CNN,

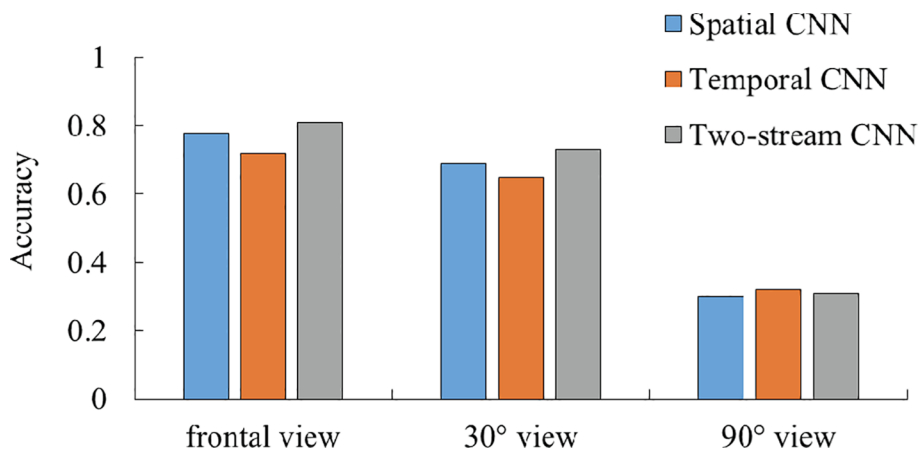

Fig. 3. Results of Simulation 2: model performance for 15-category action classification shows viewpoint-dependent effect with skeletal displays. The CNN models were trained with actions in frontal view and tested in other new views (30 and 90°away from the frontal view).

temporal CNN and two-stream CNN models. These simulation results indicate that the CNN models show a certain degree of viewpoint generalization from a trained view (e.g., frontal view) to nearby testing views (e.g., ± 30° away from frontal). Viewpoint generalization is achieved from both appearance and motion processes, manifested in the spatial CNN and temporal CNN, respectively. We used a bootstrapping approach to examine the variability of testing performance. Ten iterations of testing were performed with 20% of testing data randomly selected in each iteration. The standard deviation of testing accuracy across 10 iterations was very small (< 0.004 for all three CNNs) so error bars were not presented in figures.

However, when the testing actions were shown in 90° profile view (a large difference from the frontal view used in training), action recognition accuracy dropped significantly to ~0.3 (but still above chance level) for all three CNN models. Such viewpoint-dependent performance by CNN models is consistent with human recognition of identity and gender from walking gaits (Troje et al., 2005; Jokisch et al., 2006; Mather & Murdoch, 1994; Troje, 2002), as well as MEG results showing viewpoint-dependent representations for action recognition (Isik et al., 2017).

Together, these results indicate that both spatial and temporal pathways contribute to the generalization of action recognition performance from the trained viewpoint to nearby viewpoints (from frontal view to 30° view). However, for a large viewpoint change (from frontal view to profile view), recognition accuracy dropped significantly, although it remained above chance level. The two-stream CNN model with fusion network did not show stronger viewpoint-invariant recognition performance than did the single-stream CNNs. If viewpoint-invariant representations of biological motion relied on the later-stage representation after the integration of motion and form processing, we would expect that the fusion network could increase the generalization of recognition performance to untrained viewpoints for the two-stream CNN model. However, the present simulation result shows that adding the integration stage of appearance and motion processes did not enhance viewpoint-invariant recognition for actions. This model result supports the MEG findings (Isik et al., 2017) showing that early neural signals encode viewpoint-invariant information, rather than later stage brain activities.

5. Simulation 3: Effects of local image motion in biological motion perception

The previous two simulations have tested model performance for different display formats and viewpoints, but it remains unknown whether CNN models are able to recognize actions with noisy motion input. Beintema and Lappe (2002) found that even when local inter-frame motion signals are eliminated in point-light displays, humans are still able to recognize actions as long as stimuli preserve a certain degree of the global form revealing dynamic posture changes. In Simulation 3, we examined whether the CNN models demonstrate robust recognition performance as do humans when image motion signals are disrupted in point-light displays. If the CNN models can recognize actions in the absence of local inter-frame motion signals, we will further examine the contributions of individual pathways to action recognition. We used sequential position (SP) walkers created by Beintema and Lappe (2002) to test the CNN models, and compared model performance with human judgments in two experiments reported in their study.

As shown in Fig. 4, SP point-light walkers were generated by randomly placing points along the limbs in each frame (Beintema & Lappe, 2002). In addition to intact point-light walkers (Intact PL; Fig. 4, top), we generated eight-point SP walkers (8P), four-point SP walkers (4P), and inverted eight-point SP walkers (UD), to create the same four conditions used in Experiment 1 of the study by Beintema and Lappe. For eight-point SP walkers, eight dots were randomly positioned on the eight limb segments between joints, with one dot on each limb segment (Fig. 4, middle). In every frame of the animation, each point was reallocated to another randomly selected position on the limb. Therefore, individual dots in the SP walkers did not carry informative inter-frame motion signals reflecting the continuous trajectory of joint movements in walking actions. However, because the moving limbs constrained the possible locations for the dots, the sequence of underlying body postures was still maintained in the SP walkers (Beintema & Lappe, 2002). Similarly, the 4-point SP walker was generated by placing four points on four limbs, which were also randomly selected from the total of eight limbs in each frame (Fig. 4, bottom). The upside-down SP walker was generated by inverting the 8-point SP walker. All of the aforementioned conditions were generated from the original 98 walking actions taken from the CMU dataset.

Beintema and Lappe (2002) conducted a psychophysical experiment using a fine discrimination task on walking direction of SP walkers. In Beintena and Lappe's Experiment 2, the key experimental manipulations were to vary two stimulus parameters: the lifetime of the individual dots and the number of dots in SP walkers. With prolonged lifetime, each dot remains on the same limb position for a longer period of time and hence conveys more local image motion information. The number of dots influences how well the form of body structure can be extracted from the SP walker stimuli. With more dots along the limbs, it will be easier to perceive postures in the sequence. The experiment included 16 conditions with factorial combinations of the two stimulus parameters, so this design allows quantitative comparisons between humans and CNN models. In Simulation 3, we ran the CNN models for

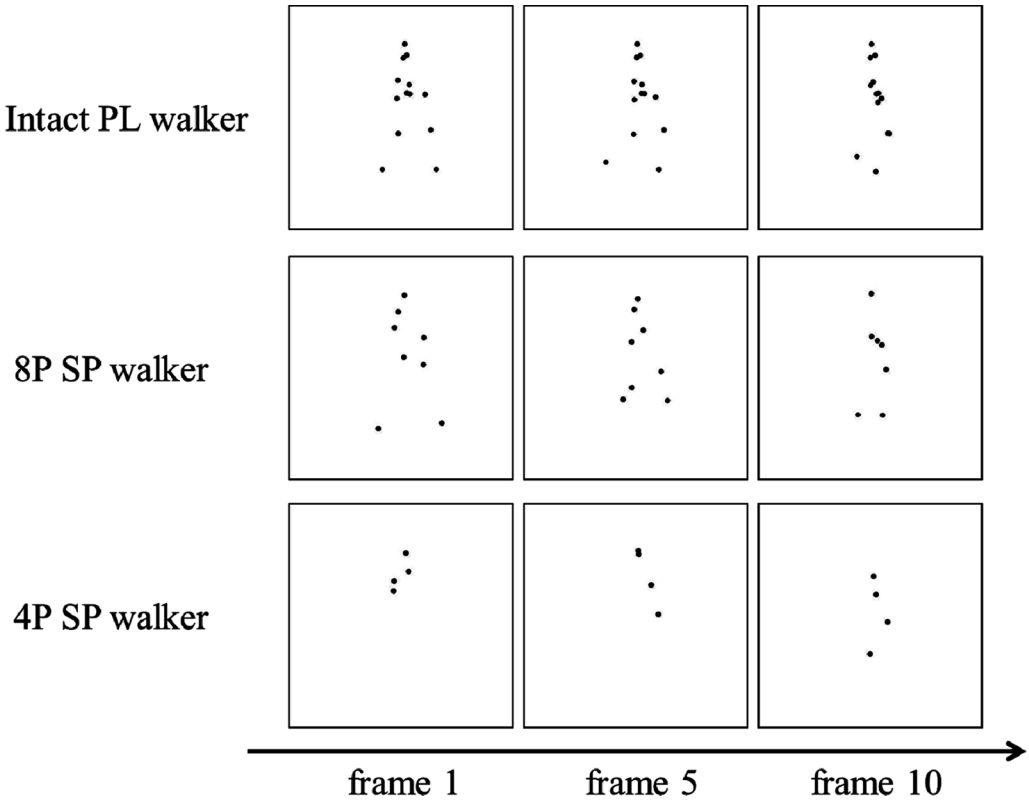

Fig. 4. Sample frames from an intact point-light (PL) walker (top), an eight-point SP walker (middle), and a four-point SP walker (bottom) in Simulation 3. Intact point-light stimuli consisted of 13 dots attached to the joints of a moving human figure. In sequential position (SP) stimuli, dots were positioned in a random location along the limbs and jumped to another randomly selected position in the next frame.

the same experimental task to examine whether the models exhibit similar performance as humans. The Supplemental section 2 includes additional simulation results for Experiment 1 in the study by Beintema and Lappe (2002).

5.1. Stimuli and model training

The walking actions in Simulation 3 were generated from the CMU motion capture database (http://mocap.cs.cmu.edu/), which includes 98 walking actions performed by 18 actors. Each walking action video lasted 2 s. Based on the motion capture data, point-light walkers were generated using the BioMotion toolbox (van Boxtel & Lu, 2013). Point-light walkers were presented in either a left profile view or a right profile view, yielding 98 instances facing left and 98 facing right. The walkers were presented in place as if walking on a treadmill. We used the same procedure as in the study of Beintema and Lappe (2002) to generate SP walkers.

Simulation 3 manipulated dot number and dot lifetime in SP walkers, generating 16 conditions. For the manipulation of dot numbers in SP walkers, each frame contained one, two, four, or eight dots (i.e., 1P, 2P, 4P, and 8P conditions). For the manipulation of dot lifetime, each dot stayed at a specific limb position for one, two, four, or eight frames before it was reallocated to another randomly selected limb position (i.e., lifetime 1, 2, 4, and 8 conditions). Initial values of lifetime were assigned randomly to each dot. Accuracy for discriminating the walking direction (left vs. right) on each trial was then measured.

First, the CNNs were trained with the 15-category action classification task by adopting the same two-step training used in Simulation 2, except that Simulation 3 used skeleton videos from the Human 3.6 M dataset with all seven viewpoints for the unrestricted transfer training. To perform the walking direction discrimination task as in the human experiment, Simulation 3 included an additional restricted transfer training for the CNN models. Specifically, the additional transfer training used 196 point-light walking actions from the CMU dataset (half leftward-facing and the other half rightward-facing) to update the connection weights in the FC and softmax layers so that the CNNs can discriminate walking directions from the PL displays (80% used for training and 20% for validation). The decision layer with 15 nodes in previous simulations was replaced by a decision layer with 2 nodes, representing leftward or rightward walking directions. After the two-step transfer training, all three CNN models achieved close-to-perfect accuracy (> 0.95) for discriminating walking direction of intact point-

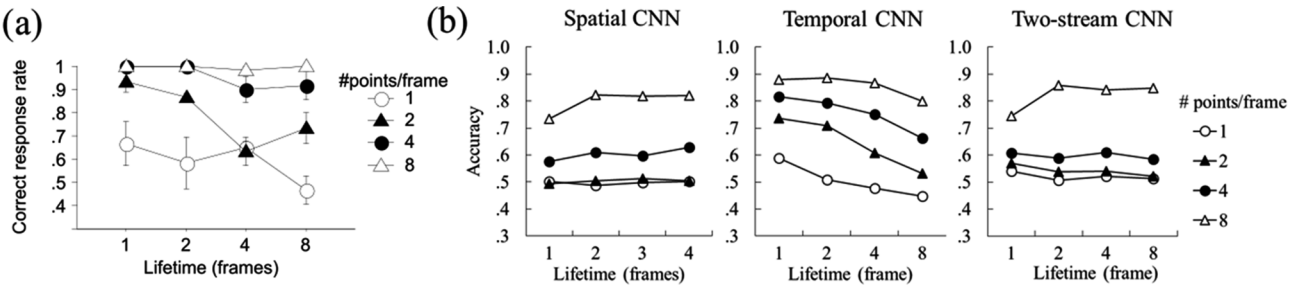

Fig. 5. Results of Simulation 3: Human and model performance for walking direction discrimination of SP walkers as a function of dot numbers and dot lifetime. (a) Human accuracy in walking direction discrimination (Beintema & Lappe, 2002, Experiment 2). Error bars represent ± 1 SE. (b) Simulation results from the spatial CNN, temporal CNN, and two-stream CNN model with fusion network.

light walkers. In Simulation 3, we used a range of SP walkers with different dot numbers and lifetime as testing stimuli.

5.2. Results and discussion

As shown in Fig. 5a, Beintema and Lappe (2002) found that human discrimination accuracy depended on the number of points in SP walkers, with better performance as more dots were included in the SP walker stimuli. However, longer lifetime of dots did not improve the discrimination performance of walking direction for SP walkers. Rather, a trend toward a slight reduction was observed as dot lifetime was increased, especially with fewer SP dots in the stimulus. This result was interpreted as indicating that fewer dots with prolonged lifetime resulted in more loss of form information in the SP stimulus.

The SP walkers were input to the spatial CNN, temporal CNN and two-stream CNN models to compute the average accuracy for the walking direction discrimination task. Results of the spatial CNN (Fig. 5b, left) revealed greater accuracy with an increasing number of dots: chance-level performance for 1P SP walkers and up to almost 0.8 accuracy for 8P SP walkers. The effect of the dot number on the performance of the spatial CNN is consistent with human behavioral patterns. Performance of the spatial CNN did not vary as a function of dot lifetime for SP walkers with 8 points and 4 points, which was consistent with the human performance for these two conditions. However, the spatial CNN failed to account for human performance with a smaller number of points in SP walker (i.e., 1 or 2 points), i.e., worse performance with prolonged lifetime. Additionally, overall recognition accuracy for the spatial CNN was lower than human accuracy. The overall correlation between human performance and predictions of the spatial CNN model was 0.70 across all 16 experimental conditions.

For the temporal CNN, the model showed an impact of the number of dots in SP walkers on discrimination accuracy, with high discrimination performance for the 8-point SP walker, and reduced performance for SP walkers with fewer dots (Fig. 5b, middle). In addition, the discrimination performance of the temporal CNN was also affected by the lifetime of dots. Especially for SP walkers with a small number of dots, model performance dropped as the lifetime of dots increased. In general, the temporal CNN qualitatively captured the impacts of dot number and lifetime on human performance observed in the psychophysical experiment with a high correlation (0.94) between model predictions and human performance. The good fit to human performance suggests that some motion features in SP walkers provided informative cues for the fine discrimination task such as walking direction discrimination. This finding is consistent with the previous work suggesting that horizontal motion in SP walkers provides reliable cues for walking direction discrimination (Casile & Giese, 2005). However, the temporal CNN predicted worse performance in 8-point SP stimuli for a prolonged lifetime, which is inconsistent with the human performance for this condition.

For the two-stream CNN model, discrimination performance overall was more similar to the spatial CNN, consistent with previous simulation results indicating that the two-stream CNN model appears to implement weighted fusion with more weight to the spatial CNN than to the temporal CNN. Across the 16 experimental conditions, the overall correlation between human performance and predictions of the two-stream CNN model was 0.69. The worse fit of the two-stream CNN than of the temporal CNN suggests that the integration in the fusion network learned from the two streams was not optimal for this specific task and SP stimuli. In contrast, humans may adjust the weighting strategy between the two streams in a more flexible way for specific stimuli and tasks.

In research on biological motion perception, the finding that people can recognize SP walkers has been used to support a template matching theory based on configural cues (Lange et al., 2006; Lange & Lappe, 2006; Theusner, de Lussanet, & Lappe, 2014). The spatial CNN could be considered as an approximation to this computational theory, as it learns a hierarchy of configural cues from a large quantity of action videos to acquire the ability to process appearance-based posture changes. The simulation results in supplemental materials (Section 2) show that the spatial CNN revealed similar activity patterns for the 8-point SP walker and the PL walker, as the template-matching theory predicts, suggesting the contribution of form processing to biological motion perception.

However, our simulation results also reveal the role of motion processing in recognizing SP walkers. What features/cues could the temporal CNN employ? It has been suggested (Casile & Giese, 2005) that SP walker stimuli contain a considerable amount of horizontal motion information that can be exploited for walking direction discrimination. The temporal CNN may take advantage of such motion cues in representing actions in SP walkers.

The interaction effect between lifetime and number of dots in human performance likely suggests that both body form and motion cues are important in supporting the recognition of biological motion. In this walking direction discrimination task, when enough dots were provided, body form provides informative cues to overcome the loss of local inter-frame motion information. However, with weakened form information due to fewer dots, prolonged lifetime of SP dots provided less informative motion cues capturing the joint movements in actions, resulting in a performance decrement. Taken together, our simulation results imply that the human visual system may integrate optical flow and body shape information overtime at different resolution levels to process the visual information in SP walkers. This integrated processing hypothesis is consistent with electrophysiological evidence that motion neurons are found in the upper bank/fundus STS of the macaque cortex and “snapshot” neurons in the lower bank of the STS and inferior temporal convexity (Vangeneugden, Pollick, & Vogels, 2009), and also with psychophysical evidence showing that local motion features (rather than global form templates) are critical for perceiving point-light biological motion (Thurman & Grossman, 2008).

6. Simulation 4: Inversion effects in biological motion perception

The inversion effect is another classic finding in biological motion perception, with people showing worse discrimination performance when point-light actions are presented upside-down (Bardi, Regolin, & Simion, 2014; Pavlova & Sokolov, 2000; Reed, Stone, Bozova, & Tanaka, 2003; Troje & Westhoff, 2006; for a review, see Blake & Shiffrar, 2007). The inversion effect has been used to support structural processing or holistic form processing in biological motion perception (Shiffrar & Pinto, 2002). However, Troje and Westhoff (2006) showed that the inversion effect can also be observed in the absence of whole-body configural information when dots in the point-light displays were spatially scrambled. This finding provided strong evidence that the human visual system is specifically tuned to some characteristic features of joint locomotion. Recent studies have found converging evidence that humans show high sensitivity to foot movements in walking actions (Wang, Zhang, He, & Jiang, 2010; Chang & Troje, 2009; van Boxtel & Lu, 2015), and to punching movements in a visual search task with boxing actions (van Boxtel & Lu, 2012). In Simulation 4, we examined whether the CNN models exhibit inversion effects after training with a large dataset of human action videos and whether these models exhibit sensitivity to critical joint movements, such as those that have been described as a “life detector” (Troje & Westhoff, 2006).

6.1. Stimuli and model training

The same 98 CMU walking actions used in Simulation 3 were used to generate test stimuli for Simulation 4. All test stimuli in Simulation 4 were created using four different types of scrambling manipulations, modeled closely on the conditions examined in Troje and Westhoff (2006) study: intact point-light, spatial scrambled, phase scrambled, and frequency scrambled conditions. *Intact point-light* displays were

generated by showing walking actions in the point-light format. *Spatially scrambled* displays were generated by randomly placing initial locations of dots of the intact point-light walker within a spatial window, but maintaining the same motion trajectory for individual dots. *Phase scrambled* displays were generated by randomizing the relative phase of dot movements (i.e., each dot started the motion sequence from a random frame in the cycle instead of starting from frame 1). *Frequency scrambled* displays were generated by scrambling the frequency of individual dot movements. The frequency scrambling manipulation was performed by multiplying the veridical speed of each dot by a ratio randomly selected within the range of 0.5 and 2 (following a uniform distribution on a logarithmic scale).

The same set of models, the spatial CNN, the temporal CNN, and the two-stream CNN trained and used in Simulation 3 were employed in Simulation 4 to perform the walking direction discrimination task for different experimental conditions.

6.2. Results and discussion

In the Troje and Westhoff (2006) study, humans were asked to judge the walking direction of intact and scrambled point-light walkers in both upright and upside-down body orientations. The researchers showed that humans were able to judge walking directions even when body configuration was disrupted by scrambling in upright orientation. Additionally, people showed clear inversion effects across all conditions, with better performance in upright orientation than in upside-down orientation for both intact point-light walkers and scrambled walkers (including spatial scramble, temporal scramble, and frequency scramble). These results indicate that the configural form of body structure is not the only cue supporting the inversion effects found in biological motion perception. Rather, local motion signals of joint movements also contribute to the well-established inversion effects.

For the intact upright point-light walkers, all three CNN models achieve 1.00 accuracy in discriminating the walking direction. As shown in Fig. 6, all three CNN models (spatial CNN, temporal CNN, and two-stream CNN model) showed inversion effects for conditions of intact point-light walker, phase scrambling, and frequency scrambling. The model performance in these three conditions are qualitatively consistent with human performance showing higher discrimination performance in the upright condition than in the upside-down condition.

However, none of the three CNN models exhibited an inversion effect in the spatial scrambling condition, in contrast with the inversion effect shown in human experiment for this condition. When dots in point-light displays were spatially scrambled, the CNN models yielded chance-level performance for direction discrimination in both upright and upside-down orientations. While phase scrambling and frequency scrambling both disturb coordinated movements of dots in the point-light display, the configural form of body structure is still preserved to a certain extent, as in these conditions the walking action is still recognizable but tends to be perceived as a wobbling walker or with an uncoordinated walking style. In contrast, spatial scrambling completely removes the configural form of the body and only preserves the pendular joint movements of a walker. Hence, the failure to discriminate walking direction in the spatially-scrambled upright walkers demonstrated that CNN models have not acquired specialized visual feature detector for biological movements (i.e., hypersensitivity to certain signature features in joint movements, such as foot movements in walking actions). These results suggest that the CNN models lack a specialized mechanism to maintain high sensitivity to critical motion of local joints (e.g., bipedal movements of feet) that signals biological movements, and/or a mechanism of passing this information directly to later layers for facilitating recognition or detection of biological movements indicative of living organisms. Another possibility of the CNN models lacking sensitivity to bipedal movements could be partially due to the smoothing procedures involved in the algorithm of calculating optical flow information from videos. The algorithm used to extract optical flow field from videos was the iterative Lucas-Kanade method with pyramids (i.e., function *calcOpticalFlowPyrLK* from the openCV toolbox). This algorithm involves smoothing image components to detect displacements over time. Especially for tracking high-speed motion, the algorithm reduces spatial resolutions of image frames. This blurring process might mitigate the precision of optical flow with fast body movements (such as foot) in action videos, yielding low sensitivity to bipedal movements in walking actions.

7. Simulation 5: Sensitivity to motion congruency in actions

In addition to recognizing actions from sparse information, humans also show the ability to perceive motion congruency in biological motion governed by causal relations (Peng et al., 2017) and to mentalize intention (Runeson & Frykholm, 1983). As a simple example, navigating the body through the environment provides humans with direct experience of cause-effect relations, because the human body moves via locomotory movements that leverage gravity and limb biomechanics to propel the body in a particular direction. This process creates a relation between limb movements as the cause and whole-body translation as the effect, resulting in expectations about the relation between the two motion cues (i.e., relative limb movements with reference to body-centered coordinates, and body displacements with reference to distal world coordinates). Several studies have shown that humans are sensitive to the congruency between relative limb movements and body displacements (Masselink & Lappe, 2015; Murphy, Brady, Fitzgerald, & Troje, 2009; Thurman & Lu, 2016; Peng et al., 2017). A compelling

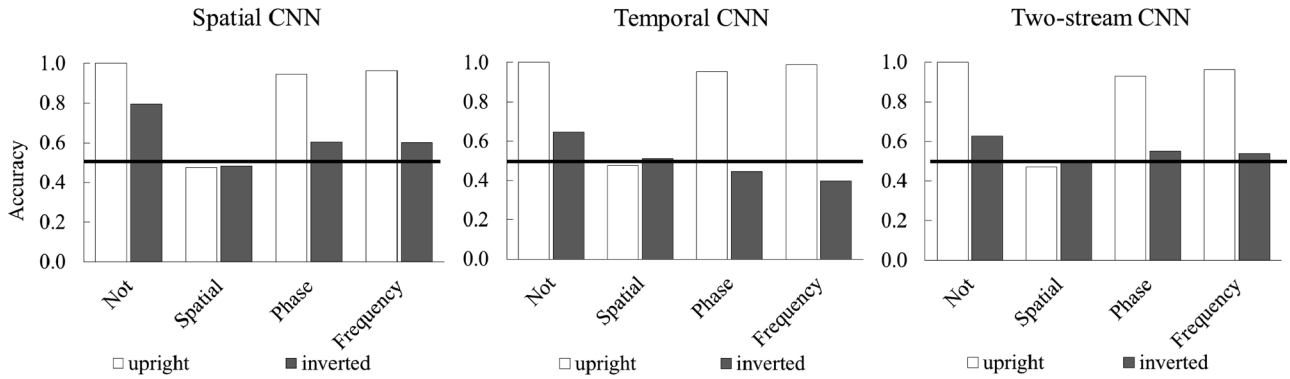

Fig. 6. Results of simulation 4: model performance for walking direction discrimination as a function of different scrambling conditions for upright and inverted walkers. Humans show an inversion effect in all four conditions (Troje & Westhoff, 2006). All CNN models showed the inversion effect in intact, phase scrambling and frequency scrambling conditions, but neither model showed the inversion effect in spatial scrambled condition.

demonstration of the strong sense of causality elicited by actions is provided by the famous “moonwalk” dance movement, which creates the illusion that the dancer is being pulled backward by an unseen force while attempting to walk forward. People seem to inevitably sense that the moonwalk movement is somehow odd and surprising.

To examine whether the two-stream CNN model supports a deeper understanding of motion information in human actions, Simulation 5 examined the ability of the CNN models to make refined discriminations among different types of walking stimuli, including intact forward walking, backward walking, moonwalk, and in-place walking. If CNNs trained for recognition of action categories fail the discrimination task, we will further test whether the CNN models can be trained to discriminate between walking actions either consistent or inconsistent with motion congruency, and whether the CNNs show sensitivity to causal directionality underlying motion congruency.

Studies of biological motion perception have shown that people are sensitive to facing direction (i.e., leftward- vs. rightward-facing) and walking directions (forward vs. backward walking) of a point-light walker (Verfaillie, 2000; Theusner et al., 2011; Lange & Lappe, 2007; Miller & Saygin, 2013; Pavlova, Krägeloh-Mann, Birbaumer, & Sokolov, 2002; Sumi, 1984; Troje & Aust, 2013). In Simulation 5 we examined a variety of walking actions, including forward/backward walking, moonwalk and walking on a treadmill. The CNN models were trained with intact walking sequences with consistent facing and walking directions, and then were tested with other walking sequences that altered facing and walking directions. If the CNN models are able to learn to be sensitive to the congruency of motion signals, then inconsistency between motion cues in the three testing action conditions would affect discrimination performance.

7.1. Stimuli and model training

The same 98 CMU walking actions employed in previous simulations were used to generate test stimuli. Some walking stimuli in Simulation 5 showed body displacements in the display. Four conditions of walking actions were generated: (1) forward walking, (2) backward walking, (3) moonwalk, and (4) in-place walking. The forward walking condition included the normal forward walking actions with consistent limb movements and body displacements, and also congruent facing direction and walking direction. The backward walking actions were generated by reversing the frame sequence of the entire video, so that limb movements and body displacements are congruent, but the walking direction is opposite to the facing direction. The moonwalk condition was generated by reversing the horizontal moving direction of the global body translation while keeping the limb movements sequence intact. Thus in a moonwalk, when a walker moves limbs in a way to naturally propel the body to move left, the body instead shifts to the right. Finally, in-place walking actions were generated by removing the global body translation component and only keeping the limb movements, as in walking on a treadmill.

Classification categories for all walking actions were defined based on the corresponding facing direction (i.e., whether the body is facing left or right regardless of limb or body movements). Fig. 7 illustrates examples that would be classified as “right” in all four conditions. As shown in Fig. 7, in the intact forward walking condition, a walker faces right and also walks towards the right. In the backward walking condition, the actor faces right, although both limb movements and body translation show leftward motion. In the moonwalk condition, the actor still faces right, and limb movements would indicate a rightward walking direction which is inconsistent with the leftward body translation shown in this condition. In the in-place walking condition, the actor faces right with the limb movements consistent with rightward motion, but the body position is stationary as in walking on a treadmill. Because simulation 1 showed that the two-stream models showed good recognition performance with skeletal displays, all the training and testing action stimuli in Simulation 5 were presented using skeletal displays. The same set of CNNs trained in Simulation 1 was used, followed by a restricted transfer training with forward walking skeleton videos to perform a discrimination task of facing directions. Training and validation stimuli were randomly chosen from 196 forward walking skeleton videos with a proportion of 80% and 20% respectively. The saturated model weights after transfer training were used for testing backward walking, moonwalk, and in-place walking conditions. Accuracy and confusion matrices were calculated for each condition.

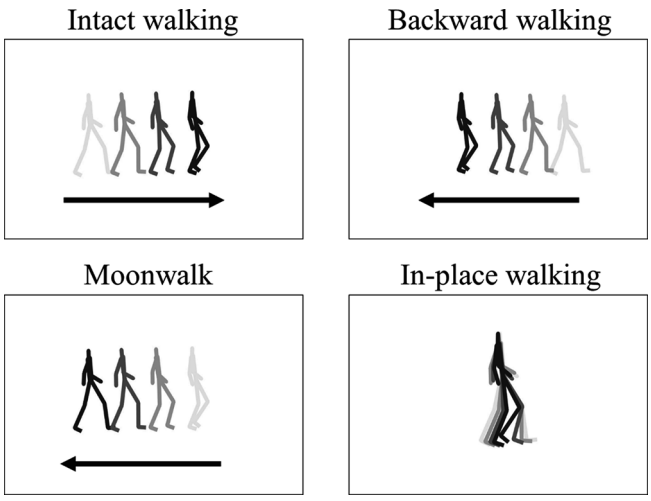

Fig. 7. Illustrations of the four walking conditions: intact forward walking, backward walking, moonwalk, and in-place walking. Here, illustrations of all conditions show instances of the “facing right” category for all four conditions. Each plot shows several possible limb movements for a stick-figure resulting from posture changes over time. The sticks in the walker change from light to dark color to denote elapsed time. Arrows below the stick-figures indicate the global body translation direction. In the in-place condition, the walker remained in a stationary location.

7.2. Results and discussion

All three CNNs showed very high performance (> 0.91) in discriminating facing directions of forward walking actors in skeletal displays. Model performance is summarized in Table 3. When testing the spatial CNN in the various walking conditions, the spatial CNN's judgment on the facing direction of walkers was influenced by the absence of body displacement, but not by the congruency of limb movements and body translation. Specifically, for in-place walking actions in the absence of body displacement, accuracy of facing direction judgments was reduced to 0.60. However, when body displacement was present in the moonwalk and backward walking conditions, the spatial CNN reached ceiling performance 0.91 of facing direction discrimination for the moonwalk condition and 0.95 for the backward walking condition, suggesting that spatial processing of appearance information extracts facing direction as long as the body moves with translation, regardless of whether the body moves to the direction in

Table 3
Training and testing response proportion to the facing direction in Simulation 5.1. From top to bottom, the rows show results for the spatial pathway, temporal pathway, and fusion.

	Training: Forward walking	Testing: Backward walking	Testing: Moonwalk	Testing: In-place walking
Appearance (spatial CNN)	0.91	0.95	0.91	0.60
Motion (temporal CNN)	0.97	0.89	0.80	0.63
Fusion (spatial + temporal two-stream CNN)	0.95	0.91	0.84	0.59

accordance with limb motion. As training action stimuli in Simulation 5 included body displacement of the sort humans usually observe human actions in the environment, it is not surprising that the spatial CNN showed sensitivity to the presence of body displacements in discrimination. However, the lack of differences between moonwalk and backward walking conditions suggests that the model has not learned the appropriate binding between posture changes from limb movements and the corresponding body translation in the environment.

The temporal CNN reached perfect accuracy (0.97) in the facing discrimination task for the forward walking actions after training. When tested with in-place walking, accuracy of the temporal CNN dropped to 0.63, revealing that the CNN likely relied on body displacement direction to decide on the facing directions of a walker. When tested with moonwalk and backward walking actions, the temporal CNN yielded performance of 0.80 and 0.89, respectively, suggesting that when body translation is opposite to the facing direction of the walker, the temporal CNN's ability in identifying the facing direction was weakened.

The two-stream CNN with fusion network showed performance intermediate between that of the spatial CNN and the temporal CNN, indicating a compromised decision based on appearance and motion features processed by the two pathways.

Simulation 5 showed that CNN models revealed small differentiation in judging the facing directions of moonwalk actions and of backward walking actions. This result may be due to the fact that these CNN models were trained to discriminate the facing direction of a walker, rather than the consistency among motion cues (in particular, motion congruency between limb movements and body displacements). In the additional simulation (see details in supplemental materials Section 3), we conducted restricted transfer training using a 3-way decision task that required the explicit differentiation among forward walking, backward walking, and moonwalk actions. Both forward walking and backward walking exhibit a causal congruency between limb movements and body translations (i.e., limb movements cause body translations), whereas moonwalk violates the causal congruency. The simulation results showed that the targeted training enables the three CNN models to acquire some sensitivity to temporal direction in accordance with the cause-effect relation in body movements.

8. General discussion

In the present study, we assessed whether single-stream CNN models and a two-stream CNN model for action recognition can account for classic findings involving human biological motion perception. Simulation 1 showed that despite attaining high accuracy after training with raw videos and skeletal displays, in comparison to humans, CNN models showed less robust performance for action recognition with novel point-light displays. Furthermore, even though the temporal CNN of motion processing produced above-chance performance, the two-stream CNN with fusion network did not show strong recognition performance for point-light stimuli, suggesting that the integration stage in the two-stream CNN overweights the image features extracted by the spatial CNN based on appearance processing, but underweights the motion features from the temporal CNN.

In Simulation 2, CNN models showed viewpoint-dependent recognition and limited ability to generalize from a trained viewpoint to nearby views. In Simulation 3, we found that the CNN models showed some ability to recognize walking actions when local image motion signals are disrupted in SP walkers. Both the spatial CNN based on appearance processing and the temporal CNN based on motion processing contribute to the recognition of walkers with degraded motion information. Simulation 4 revealed that the CNN models predict the inversion effect attributable to global configural cues, but fail to predict the inversion effect attributable to specialized local motion cues (i.e., “life detectors”).

Simulation 5 systematically examined whether CNN models can capture more fine-grained features of action stimuli, such as causal congruency between motion cues. Simulation 5 trained the CNNs with a facing-direction discrimination task. We found that CNNs demonstrated a certain degree of sensitivity to the presence of global body displacement in action stimuli, as the models showed worse performance for in-place walkers. However, the CNNs did not show clear differentiation for backward walking and moonwalk. Additional simulation used a targeted task to train the CNNs to discriminate forward walking, backward walking and moonwalk. After training, all three CNN models showed some sensitivity to temporal direction in accordance with the cause-effect relation in body movements.

Together, these findings indicate that CNN models can achieve near human-level performance in action recognition after intensive training with raw RGB videos, and show a certain degree of generalization to novel displays of biological motion after transfer training. However, the CNN models have not achieved the robustness as human perception of biological motion. In particular, CNNs trained with raw RGB videos show weak performance in recognizing point-light actions, which contrasts with humans’ remarkable ability to perform point-light action recognition without any need for extra training. In object recognition, researchers found that CNN models primarily rely on the statistical regularities of low-level appearance information to perform visual recognition, and lack the ability to extract global shape cues (Baker et al., 2018). The CNN models for action recognition exhibit the similar weakness, showing limited generalization from training data (raw videos) to other display types (e.g., point-light display). Additional transfer training is necessary for the CNN models to recognize actions in point-light displays. Whereas the CNN models rely heavily on a large sample of training instances, the human visual system can form a concept from even a single encounter (Carey & Bartlett, 1978). Even though CNN models are powerful in capturing the appearance patterns and motion kinematics from action videos, they lack a high-level abstract representation of actions.

Another shortcoming of the two-stream CNN model involves the integration module within the fusion network, which appears to assign a higher weight to the spatial stream of processing appearance than to the temporal stream of processing motion after training with raw RGB videos. In four simulations (i.e., simulations 1 – 4), the two-stream CNN with fusion network showed similar performance as that of the spatial CNN. This fusion strategy may be optimal for the trained task and dataset with raw RGB videos. However, the lack of flexibility in adjusting the weighting strategy between form processing and motion processing significantly limits the model’s ability to achieve human-level generalization to novel stimuli.

All three CNN models lack sensitivity to specialized motion cues that signal animacy or life in biological motion. As has been shown in studies with adult humans (Troje & Westhoff, 2006), newborns (Bidet-Ildei, Kitromilides, Orliaguet, Pavlova, & Gentaz, 2014), and newly-hatched chicks (Vallortigara & Regolin, 2006), characteristic movements of feet serve as an important indicator of living animals in locomotion, and attract visual attention automatically. The two-stream CNN model does not have a mechanism to differentiate visual filters that are tuned to specialized movement patterns such as a life detector. Furthermore, due to the architecture of the CNN models, the lack of long-range connections across layers makes it difficult to directly pass certain critical local motion cues to later decision layers in support of efficient detection of biological motion. Whereas the human visual system may be able to detect life signals based on scattered motion signals and flexibly assemble motion information by integrating different visual cues (Thurman & Lu, 2013) to form biological motion representations for novel creatures, CNN models have limited ability to adaptively integrate local motion information and global body form for specific tasks and stimuli.

Finally, many psychophysical studies have revealed important top-down influences on biological motion perception (Lu, Tjan, & Liu, 2006), an interplay with motor training (Casile & Giese, 2006) and

social perception (Johnson, McKay, & Pollick, 2011; Klin, Lin, Gorrindo, Ramsay, & Jones, 2009), and interaction with visual attention (Thornton & Vuong, 2004; Thompson & Parasuraman, 2012). In contrast, the CNN models are constructed in a pure feedforward manner for both spatial and temporal pathways, without top-down influences through feedback connections and interactions between the two pathways. This architecture enables the model to learn visual patterns (appearance and motion) associated with action categories, but limits its capability to manipulate attention and to incorporate prior knowledge, such as physical laws and biological constraints.

These shortcomings indicate that CNN models for action recognition are susceptible to a mismatch between training and testing datasets, due to their limited ability to form robust representations of human body movements. Controlled stimuli commonly used in psychophysical studies provide a useful tool to assess the generalization ability of CNNs. Future work should focus on overcoming the aforementioned limitations to enhance the models' generalizability to novel stimuli and a larger range of tasks, rather than focusing solely on recognition accuracy for a specific task. In addition to the behavioral findings explored in the present paper, many other psychophysical effects in biological motion perception (e.g., size invariance, embodiment) can be used to further gauge the underlying representational commonalities and differences between human action perception system and the operation of CNN models.

CRediT authorship contribution statement

Yujia Peng: Conceptualization, Methodology, Data curation, Formal analysis, Software, Investigation, Visualization, Validation, Writing - original draft, Project administration. **Hannah Lee:** Formal analysis, Investigation, Data curation, Software, Writing - review & editing. **Tianmin Shu:** Methodology, Software, Validation, Writing - review & editing. **Hongjing Lu:** Conceptualization, Supervision, Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by a CSC scholarship to Peng and NSF grant BSC-1655300 to Lu. We thank Qi Xie for assistance in running simulations, Chisei Mizuno for help with data preparation, and Joseph M. Burling for aid in setting up the model. We thank Prof. Lappe for sharing the human experiment data for Simulation 3.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.visres.2020.09.005.

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology, 14*(12), Article e1006613.

Bardi, L., Regolin, L., & Simion, F. (2014). The first time ever I saw your feet: Inversion effect in newborns' sensitivity to biological motion. *Developmental Psychology, 50*(4), 986.

Beintema, J. A., & Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences, USA, 99*(8), 5661–5663.

Bidet-Ildei, C., Kitromilides, E., Orliaguet, J. P., Pavlova, M., & Gentaz, E. (2014). Preference for point-light human biological motion in newborns: Contribution of translational displacement. *Developmental Psychology, 50*(1), 113.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology, 58*, 47–73.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience, 16*(11), 3737–3744.

Brownlow, S., Dixon, A. R., Egbert, C. A., & Radcliffe, R. D. (1997). Perception of movement and dancer characteristics from point-light displays of dance. *Psychological Record, 47*(3), 411–422.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17–29.

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision, 5*(4), 6.

Casile, A., & Giese, M. A. (2006). Nonvisual motor training influences biological motion perception. *Current Biology, 16*(1), 69–74.

Chang, D. H., & Troje, N. F. (2009). Acceleration carries the local inversion effect in biological motion perception. *Journal of Vision, 9*(1), 19.

Cutting, J. E., & Kozlowski, L. T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society, 9*(5), 353–356.

Cutting, J. E., Moore, C., & Morrison, R. (1988). Masking the motions of human gait. *Perception & Psychophysics, 44*(4), 339–347.

Dittrich, W. H., Troscianko, T., Lea, S. E., & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception, 25*(6), 727–738.

Dow, B. M., Snyder, A. Z., Vautin, R. G., & Bauer, R. (1981). Magnification factor and receptive field size in foveal striate cortex of the monkey. *Experimental Brain Research, 44*, 213–228.

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1933–1941). .

Gallant, J. L., Braun, J., & Van Essen, D. C. (1993). Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science, 259*(5091), 100–103.

Giese, M. A., & Poggio, T. (2003). Cognitive neuroscience: Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*(3), 179.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron, 35*(6), 1167–1175.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience, 12*(5), 711–720.

Hegdé, J., & Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience, 20*(5), RC61.

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17*(1–3), 185–203.

Ionescu, C., Li, F., & Sminchisescu, C. (2011). Latent structured models for human pose estimation. *2011 IEEE International Conference on Computer Vision* (pp. 2220–2227). IEEE.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(7), 1325–1339.

Isik, L., Tacchetti, A., & Poggio, T. (2017). A fast, invariant representation for human action in the visual system. *Journal of Neurophysiology, 119*(2), 631–640.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*(2), 201–211.

Johnson, K. L., McKay, L. S., & Pollick, F. E. (2011). He throws like a girl (but only when he's sad): Emotion affects sex-decoding of biological motion displays. *Cognition, 119*(2), 265–280.

Jokisch, D., Daum, I., & Troje, N. F. (2006). Self recognition versus recognition of others by biological motion: Viewpoint-dependent effects. *Perception, 35*(7), 911–920.

Klin, A., Lin, D. J., Gorrindo, P., Ramsay, G., & Jones, W. (2009). Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature, 459*(7244), 257–261.

Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics, 21*(6), 575–580.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1097–1105). .

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). *HMDB: a large video database for human motion recognition* (pp. 2556–2563). IEEE.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *Journal of Neuroscience, 26*(11), 2894–2906.

Lange, J., Georg, K., & Lappe, M. (2006). Visual perception of biological motion by form: A template-matching analysis. *Journal of Vision, 6*, 836–849.

Lange, J., & Lappe, M. (2007). The role of spatial and temporal information in biological motion perception. *Advances in Cognitive Psychology, 3*(4), 419.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5*(5), 552–563.

Lu, H. (2010). Structural processing in biological motion perception. *Journal of Vision, 10*(12), 1–13.

Lu, H., Tjan, B. S., & Liu, Z. (2006). Shape recognition alters sensitivity in stereoscopic depth discrimination. *Journal of Vision, 6*(1), 7.

Masselink, J., & Lappe, M. (2015). Translation and articulation in biological motion perception. *Journal of vision, 15*(11) 10–10.

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 258*(1353), 273–279.

Mather, G., Radford, K., & West, S. (1992). Low-level visual processing of biological motion. *Proceedings of the Royal Society of London. Series B: Biological Sciences,*

249(1325), 149–155.
Miller, L. E., & Saygin, A. P. (2013). Individual differences in the perception of biological motion: links to social cognition and motor imagery. *Cognition, 128*(2), 140–148.
Murphy, P., Brady, N., Fitzgerald, M., & Troje, N. F. (2009). No evidence for impaired perception of biological motion in adults with autistic spectrum disorders. *Neuropsychologia, 47*(14), 3225–3235.
Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. *Nature, 395*(6705), 894.
Oram, M. W., & Perrett, D. I. (1994). Responses of anterior superior temporal polysensory (STPa) neurons to "biological motion" stimuli. *Journal of Cognitive Neuroscience, 6*(2), 99–116.
Pavlova, M., Krägeloh-Mann, I., Birbaumer, N., & Sokolov, A. (2002). Biological motion shown backwards: the apparent-facing effect. *Perception, 31*(4), 435–443.
Pavlova, M., & Sokolov, A. (2000). Orientation specificity in biological motion perception. *Perception & Psychophysics, 62*(5), 889–899.
Pavlova, M. A. (2011). Biological motion processing as a hallmark of social cognition. *Cerebral Cortex, 22*(5), 981–995.
Peng, Y., Thurman, S., & Lu, H. (2017). Causal action: A fundamental constraint on perception and inference about body movements. *Psychological Science, 28*(6), 798–807.
Pinto, J., & Shiffrar, M. (1999). Subconfigurations of the human form in the perception of biological motion displays. *Acta Psychologica, 102*(2–3), 293–318.
Pollick, F. E., Kay, J. W., Heim, K., & Stringer, R. (2005). Gender recognition from point-light walkers. *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1247.
Reed, C. L., Stone, V. E., Bozova, S., & Tanaka, J. (2003). The body-inversion effect. *Psychological Science, 14*(4), 302–308.
Rodman, H. R., & Albright, T. D. (1989). Single-unit analysis of pattern-motion selective properties in the middle temporal visual area (MT). *Experimental Brain Research, 75*(1), 53–64.
Runeson, S., & Frykholm, G. (1983). Kinematic specification of dynamics as an informational basis for person-and-action perception: expectation, gender recognition, and deceptive intention. *Journal of experimental psychology: general, 112*(4), 585.
Shiffrar, M., & Pinto, J. (2002). Are we visual animals? *Journal of Vision, 2*(7) 334–334.
Simonyan, K., & Zisserman, A. (2014). *Two-stream convolutional networks for action recognition in videos. Proceedings of neural information processing systems*.
Smith, A. T., & Snowden, R. J. (Eds.). (1994). *Visual detection of motion*. New York: Academic Press.
Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.
Sumi, S. (1984). Upside-down presentation of the Johansson moving light-spot pattern. *Perception, 13*(3), 283–286.
Theusner, S., de Lussanet, M. H., & Lappe, M. (2011). Adaptation to biological motion leads to a motion and a form aftereffect. *Attention, Perception, & Psychophysics, 73*(6), 1843–1855.
Theusner, S., de Lussanet, M., & Lappe, M. (2014). Action recognition by motion detection in posture space. *Journal of Neuroscience, 34*(3), 909–921.
Thompson, J., & Parasuraman, R. (2012). Attention, biological motion, and action recognition. *Neuroimage, 59*(1), 4–13.
Thornton, I. M., & Vuong, Q. C. (2004). Incidental processing of biological motion. *Current Biology, 14*(12), 1084–1089.
Thurman, S. M., & Grossman, E. D. (2008). Temporal "bubbles" reveal key features for point-light biological motion perception. *Journal of Vision, 8*(3), 28.
Thurman, S. M., & Lu, H. (2013). Complex interactions between spatial, orientation, and motion cues for biological motion perception across visual space. *Journal of Vision, 13*(2), 8.
Thurman, S. M., & Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PLoS ONE, 9*(11), Article e112539.
Thurman, S. M., & Lu, H. (2016). A comparison of form processing involved in the perception of biological and nonbiological movements. *Journal of vision, 16*(1) 1–1.
Thurman, S. M., van Boxtel, J. J., Monti, M. M., Chiang, J. N., & Lu, H. (2016). Neural adaptation in pSTS correlates with perceptual aftereffects to biological motion and with autistic traits. *Neuroimage, 136*, 149–161.
Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision, 2*(5), 2.
Troje, N. F., & Aust, U. (2013). What do you mean with "direction"? Local and global cues to biological motion perception in pigeons. *Vision Research, 79*, 47–55.
Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a "life detector"? *Current Biology, 16*(8), 821–824.
Troje, N. F., Westhoff, C., & Lavrov, M. (2005). Person identification from biological motion: Effects of structural and kinematic cues. *Perception & Psychophysics, 67*(4), 667–675.
Vaina, L. M., Lemay, M., Bienfang, D. C., Choi, A. Y., & Nakayama, K. (1990). Intact "biological motion" and "structure from motion" perception in a patient with impaired motion mechanisms: A case study. *Visual Neuroscience, 5*(4), 353–369.
Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., & Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences, USA, 98*(20), 11656–11661.
Vallortigara, G., & Regolin, L. (2006). Gravity bias in the interpretation of biological motion by inexperienced chicks. *Current Biology, 16*(8), R279–R280.
van Boxtel, J. J., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. *Journal of Vision, 13*(12), 7.
van Boxtel, J. J., & Lu, H. (2015). Joints and their relations as critical features in action discrimination: Evidence from a classification image method. *Journal of Vision, 15*(1), 20.
Vangeneugden, J., Pollick, F., & Vogels, R. (2009). Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. *Cerebral Cortex, 19*(3), 593–611.
Verfaillie, K. (2000). Perceiving human locomotion: Priming effects in direction discrimination. *Brain and Cognition, 44*(2), 192–213.
Wang, L., Zhang, K., He, S., & Jiang, Y. (2010). Searching for life motion signals: Visual search asymmetry in local but not global biological-motion processing. *Psychological Science, 21*(8), 1083–1089.